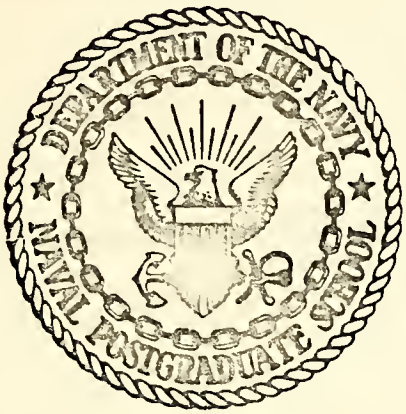
SOME ALTERNATIVE METHODS TO STEPWISE REGRESSION FOR THE SCREENING OF VARIABLES

Dennis George Lambell

NAVAL POSTGRADUATE SCHOOL
Monterey, California



THESIS

SOME ALTERNATIVE METHODS TO STEPWISE REGRESSION FOR THE SCREENING OF VARIABLES

by

Dennis George Lambell

December 1974

Thesis Advisor: R. R. Read

Approved for public release; distribution unlimited.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO. / 3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): Some Alternative Methods to Stepwise Regression for the Screening of Variables	5. TYPE OF REPORT & PERIOD COVERED: Master's Thesis; December 1974
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): Dennis George Lambell	8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, California 93940	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Postgraduate School, Monterey, California 93940	12. REPORT DATE: December 1974
	13. NUMBER OF PAGES: 54
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Naval Postgraduate School, Monterey, California 93940	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Principle Component Analysis
Stepwise Regression
Total Enumeration
Multiple Correlation Coefficient

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper addresses the problem of screening potential variables for entrance in a linear multiple regression setting. The purpose of the work presented here is to propose two screening methods, both of which have roots in principle component analysis, and which evaluate a combination of variables in an efficient enough manner so that enumeration of all combinations is feasible even when the number of

Block #20 Continued

potential variables is quite large. Using the square of the multiple correlation coefficient as the criterion, the selections made by these methods in several test cases are evaluated, and compared with the selections made by the methods of total enumeration and stepwise regression. The paper concludes with overall evaluations of the two methods and suggests directions for further study.

Some Alternative Methods to Stepwise Regression for the Screening of Variables

by

Dennis George Lambell
Ensign, United States Navy
B.S., California State University at Los Angeles, 1973

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN OPERATIONS RESEARCH

from the

NAVAL POSTGRADUATE SCHOOL
December 1974

ABSTRACT

This paper addresses the problem of screening potential variables for entrance in a linear multiple regression setting. The purpose of the work presented here is to propose two screening methods, both of which have roots in principle component analysis, and which evaluate a combination of variables in an efficient enough manner so that enumeration of all combinations is feasible even when the number of potential variables is quite large. Using the square of the multiple correlation coefficient as the criterion, the selections made by these methods in several test cases are evaluated, and compared with the selections made by the methods of total enumeration and stepwise regression. The paper concludes with overall evaluations of the two methods and suggests directions for further study.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

The author wishes to express his appreciation to Professor Robert R. Read, for his patience, encouragement, and counsel, and for his overall contribution to this thesis effort.

I. INTRODUCTION

The search for the important variables in a multiple linear regression setting is, due to its great practical importance, an area which has received considerable attention. Researchers often use a large number of potentially important variables in the exploratory stages of their work. These need be screened in order to determine the most parsimonious subset of these variables available for predicting or estimating the response with an acceptable level of error. It appears that a total enumeration of all combinations of variables [Ref. 1] is necessary to be assured of making the best selection. This process is computationally overwhelming when the number of variables becomes even moderately large. (Each combination requires a matrix inversion, and there are 2^p such inversions to perform if p is the number of variables.)

A highly popular approach to this problem is the use of stepwise regression [Refs. 2, 3, and 4]. It is basically a one-step look-ahead method, and uses significance tests based on distributional assumptions to judge the combinations of variables under consideration at each step. It appears to do an adequate job of selection, and is readily available in packaged form (specifically BMD and SPSS).

The purpose of the present work is to present and study some alternatives to stepwise regression in hopes of finding viable competitors. It is desirable that these methods

should perform at least as well as stepwise regression, yet remain feasible computationally. In this regard the method of principle component analysis of the antecedent variables is useful and serves as a guide.

This paper will pursue the development of these alternatives in the following manner. A discussion of the general problem will be presented first, along with remarks concerning ways of measuring the effectiveness of the combinations of variables. Comments concerning several suggested screening methods will follow, being followed in turn by a discussion of stepwise regression. A development of the alternatives under consideration in this paper will be presented, and a comparative evaluation of their performance on some test cases will conclude the work.

II. THE LINEAR REGRESSION MODEL

In order to introduce the linear model, it is convenient to agree to a common notation. Let y be the dependent, or response, variable, and $x_1, x_2, \ldots, x_p$ be the control, or antecedent, variables. Assume that N sets $(y, x_1, x_2, \ldots, x_p)$ are observed, and for convenience, let each member of the set be replaced with its deviation from the sample mean:

$$y_j \leftarrow (y_j - \bar{y}) \text{ and } x_{ij} \leftarrow (x_{ij} - \bar{x}_i) \quad i=1,\ldots,p; \; j=1,\ldots,N. \tag{2.1}$$

The response y is, of course, viewed as random. The antecedent variables may be either random or deterministic; it does not matter which. We are concerned only with the question of which subset of them should be permanently collected and not with formal statistical inference *per se*. When means, variances, covariances, and correlations are introduced, they refer only to the sample quantities. Further, no distributional assumptions are made about y. Thus decisions regarding the appropriateness of the various combinations of variables are structured on *ad hoc* data analysis grounds and not on formal tests.

Using the column vector Y to denote the $\{y_j\}_1^N$, and the matrix X for the N sets of $\{(x_{1j}, \ldots, x_{pj})\}_1^N$, the usual linear model

$$\underset{N\times 1}{Y} = \underset{N\times p}{X} \; \underset{p\times 1}{\beta} + \underset{N\times 1}{\varepsilon} \tag{2.2}$$

is assumed, where β represents the vector of regression coefficients and ε the vector of residuals. The estimation of

β is achieved by the method of least squares, the solution being [Refs. 2 and 5]:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.3}$$

assuming X has full rank (as it generally will in data analysis situations) and the prime denotes transpose. The correlation matrix of the $x_1,\ldots,x_p$ variables is introduced as

$$C_{pxp} = \frac{1}{N} X'X. \tag{2.4}$$

The computational problem associated with total enumeration can now be made more explicit. Consider a subset of size q, $(x_{i_1},\ldots,x_{i_q})$ of $(x_1,\ldots,x_p)$. There are $\binom{p}{q}$ such subsets, and for each of these a covariance matrix (a minor of (2.4)) must be inverted to produce the corresponding $\hat{\beta}$ (Eqn. (2.3)). This done, one must choose the best (in some sense) subset of size q and do this for each q=1,...,p.

A number of criteria for judging the fit of a subset of variables are available, including multiple correlation, standardized total squared error [Ref. 6], and variance of residuals. The approach taken here is to chart the growth of the square of the multiple correlation coefficient, R^2, as a function of q. This can be done for total enumeration (using that subset $(x_{i_1},\ldots,x_{i_q})$ that maximizes R^2 for each q), for stepwise regression, and the two alternatives to be introduced. Once such charts are made, the user may choose q to meet his own needs. The researcher hopes that R^2 grows very rapidly and becomes very close to its maximum (achieved

when q=p) for small q. The worse case occurs when R^2 grows linearly with q.

This approach seems simple and reasonable. No formal significance tests are made and the tenant difficulties connected with simultaneous inference are not addressed. Although the method of stepwise regression uses formal significance testing in its intermediate stages, the results of using it can still be compared using our simple *ad hoc* approach.

Remark: In recent work with the method of ridge regression [Ref. 7] the use of unbiased estimators Eqn. (2.3) is foregone and more general measures based on average squared error are used. In this method the principle diagonal of the covariance matrix Eqn. (2.4) is loaded in an effort to trade bias against a smaller mean squared error. Comparison of the methods presented here with those of ridge regression is not considered in this thesis.

To present computational formulae for the measures of effectiveness discussed previously, additional notation is convenient. Let s be the pxl column vector of covariances between y and $(x_1,\ldots,x_p)$, and let c_{yy} be the variance of y. The variance of residulas can be estimated as

$$\hat{\sigma}^2 = \frac{1}{N}\,\hat{\varepsilon}\cdot\hat{\varepsilon} = \frac{1}{N}\,(Y-X\hat{\beta})'\cdot(Y-X\hat{\beta}) \qquad (2.5)$$

where

$$\hat{\varepsilon} = Y-X\hat{\beta} \qquad (2.6)$$

is the estimate of residuals. The square of the multiple correlation coefficient is [Ref. 5]

$$R^2 = 1 - \hat{\sigma}^2 / c_{yy} = s'C^{-1}s / c_{yy} \qquad (2.7)$$

and C is given in Eqn. (2.4). When the model is reduced to $(x_{i_1},\ldots,x_{i_q})$, the matrices and vectors must be modified accordingly.

Three test cases are introduced to evaluate the methods under consideration (Tables I-III). There the pertinent quantities C, s, and c_{yy} are presented as augmented matrices in the format

$$\left[\begin{array}{c|c} c_{yy} & s' \\ \hline s & C \end{array}\right] .$$

The first test case was generated specifically to expose a weakness of the stepwise approach, as will be seen. The second matrix was designed to be of sufficient complexity to give a more enlightening comparison of the methods under consideration, yet small enough so that the total enumeration solution could be obtained. Care was taken to ensure that the criterion of positive definiteness (see Appendix B) was met.

Because the first two examples are artificial in nature, an application using real data was sought. Such was obtained from a study currently underway [Ref. 8]. There, a survey concerning subjective reactions to a set of fourteen drugs is being made to ascertain how the subjects perceive the various drugs. Each drug is rated from one to seven on each of the following fourteen scales; violence, growth, sharpness, destruction, chancement, activity, goodness,

avoidance, integration, positivity, permanence, speed, severity, and strength.

One of the studies within this investigation is to describe the scale "severity" in terms of the other scales. More specifically, what subset of the other scales "best" describes severity? This problem presents an opportunity to compare the screening methods being presented here with stepwise regression, and provides a third test matrix. This data is rounded to one digit to conserve space, and in this form may not be positive definite. See Table III. Of course, the original matrix was used in the computations.

1.0	0.7	0.6	0.6
0.7	1.0	0.6	0.6
0.6	0.6	1.0	0.1
0.6	0.6	0.1	1.0

Table I
TEST MATRIX ONE

1.0	0.6	0.1	0.4	0.3	0.5
0.6	1.0	0.4	0.1	0.2	0.5
0.1	0.4	1.0	0.3	0.4	0.2
0.4	0.1	0.3	1.0	0.2	0.4
0.3	0.2	0.4	0.2	1.0	0.3
0.5	0.5	0.2	0.4	0.3	1.0

Table II
TEST MATRIX TWO

1.0	-.2	0.1	-.4	0.4	-.2	0.4	-.3	0.3	-.2	-.3	0.2	-.4	0.2
-.2	1.0	0.4	0.5	-.3	-.1	-.4	0.4	-.3	0.4	0.3	0.1	0.4	0.0
0.1	0.4	1.0	0.3	-.3	-.4	-.2	0.3	-.2	0.4	0.1	0.4	0.1	0.2
-.4	0.5	0.3	1.0	-.4	0.0	-.6	0.6	-.4	0.5	0.4	0.0	0.5	-.1
0.3	-.3	-.3	-.4	1.0	0.3	0.5	-.4	0.5	-.4	0.2	-.1	-.3	0.1
-.2	-.1	-.4	0.0	0.3	1.0	0.1	0.0	0.1	-.1	0.1	-.3	0.0	-.2
0.4	-.4	-.2	-.6	0.5	0.1	1.0	-.7	0.5	-.6	-.3	0.0	-.5	0.1
-.3	0.4	0.3	0.6	-.4	-.1	-.7	1.0	-.5	0.6	0.3	0.0	0.5	-.1
0.3	-.3	-.2	-.4	0.4	0.1	0.5	-.5	1.0	-.5	-.3	0.0	-.4	0.1
-.2	0.4	0.3	0.5	-.4	-.1	-.6	0.6	-.5	1.0	0.3	0.1	0.4	0.0
-.3	0.3	0.1	0.4	-.2	0.1	-.3	0.3	-.3	0.3	1.0	-.1	0.5	-.1
0.2	0.1	0.4	0.0	-.1	-.3	0.0	0.0	0.0	0.0	-.1	1.0	0.0	0.4
-.4	0.4	0.1	0.5	-.3	0.0	-.5	0.5	-.4	0.4	0.5	0.0	1.0	-.2
0.2	0.0	0.2	-.1	0.1	-.2	0.1	-.1	0.1	0.0	-.1	0.4	-.2	1.0

Table III
TEST MATRIX THREE
(Entries rounded; see Ref. 8 for usable data)

III. CURRENTLY USED SCREENING PROCEDURES

There are a number of methods currently used to screen variables, among them forward selection, backwards elimination, stepwise regression, and several graphical techniques [Ref. 6]. Because stepwise regression incorporates the best of two of the above methods, enjoys general acceptance, and is readily available as a packaged program, it will serve as a baseline for measuring the performance of the methods under study here, and will be discussed in greater detail.

Stepwise regression is based on an underlying assumption of normality for the response variable y. As a result of this, sums of squares from several sources (including residual error, regression, and reduced model) have Chi-Square distributions. Thus at any step, a potential new variable may be tested for its significance if allowed to enter the system. Among those eligible to enter, the most significant (in terms of the F statistics being formed) is selected. Then those variables entered at previous steps are tested to determine whether their presence remains significant. Again, statistics are used as the criteria for deletion. When no variables can either enter or leave, the process terminates. Stepwise regression has the ability to look ahead only one variable at a time, and thus cannot guarantee an optimum solution.

One of the most curious characteristics of stepwise regression as it is used in the packaged programs available

(specifically the BMD and SPSS regression packages) is the set of critical points for the F-statistic used by the computer as criteria for entrance of variables. In order to reduce core requirements these packaged programs allow a single value from the F table to be used as the criterion, despite the change in degrees of freedom required each time a new F-statistic is formed. Thus what may be a test of significance at the level α for one F-statistic with p and q degrees of freedom will not remain so for a new F-statistic with r and s degrees of freedom. As a result, there is no easy way to control the actual level of significance of each test performed (much less the overall level of significance of the final combination chosen with the multiple tests). Further, the default values (for entering a variable) in both the SPSS and BMD regression packages are set at $F_{critical}=0.01$. While these may be changed by the user, many users are unaware of the problem and rely on the default value. This default value is not one which most users would initially choose, as for most F tests this would correspond to a significance level close to one. (In fairness to stepwise regression it should be stated that the problem is with the packaged programs, not with the method itself.)

Two advantages to stepwise regression are worth noting. The first is that it only needs to look at a small subset of the total number of combinations before terminating, and therefore relatively large problems become computationally feasible. The second is that when the underlying assumptions are met the results of its screening seem to be generally quite good.

There are a number of disadvantages to stepwise regression. The first is that it is extremely difficult to "see into" the method and understand what it is doing at any step. The computer printout is confusing, and many users may be only dimly aware of what is happening. Consequently, the results obtained may often be unsatisfactory, as the user delegates too much analysis to the computer program.

A second disadvantage is that of the underlying distributional assumptions. Robustness to departures from the normality assumption becomes an issue, as well as the previously stated problem of simultaneous inference.

IV. SCREENING OF VARIABLES

The method of principle components [Refs. 2 and 5] applied to the antecedent variables provides a platform for introducing the screening methods presented in this paper. Further, the growth of the multiple correlation coefficient when the principle components are used as antecedent variables is easily developed and serves as a rough standard for the kind of growth that may be available using the original variables.

The mathematical structure is useful in that the component variables are orthogonal (uncorrelated), and their variances are readily obtainable, as are their correlations with the response variable. The matrix of eigenvectors serves to expose those antecedent variables which exert most influence over the orientation of the original data. Thus it seems reasonable that useful screening methods can be obtained by first finding the important principle components, and then the important original variables that influence these components.

General developments of the method of principle components are readily available (see for example [Refs. 2 and 5]). The salient properties used herein are presented below and, for sake of immediate reference, developed in Appendix A.

The rotation of the vector of antecedent variables x to the principle components vector v may be expressed as follows:

$$v = W'x \tag{4.1}$$

where the columns of W, $(w_1, w_2, \ldots, w_p)$, are the eigenvectors of C (Eqn. (2.4)). The covariance matrix of v is diagonal with the variances being the eigenvalues. The covariance of a typical v_i with the response y can be calculated as

$$\mathrm{Cov}(y, v_i) = E(yv_i) = w_i' E(yx) = w_i' s \tag{4.2}$$

and hence the correlations are

$$r_{y,v_i} = \frac{w_i' s}{\sqrt{\lambda_i c_{yy}}} = \frac{1}{\sqrt{\lambda_i}} \sum_{j=1}^{p} w_{ji} \sqrt{C_{jj}}\ r_{y,x_j} \tag{4.3}$$

where w_{ij} are the elements of W and C_{ij} the elements of C.

When the principle components are considered as being the antecedent variables, it is an easy task to chart the square of the multiple correlation coefficient R^2 as a function of the number of variables q. One need only order the $(v_1, \ldots, v_p)$ according to the magnitudes of their correlations with y (given by Eqn. (4.3)):

$$r^2_{y,v_1} > r_{y,v_2} > \cdots > r^2_{y,v_p}. \tag{4.4}$$

It follows from (2.7) that

$$\begin{aligned} R^2_{pc} &= \sum_{i=1}^{q} r^2_{y,v_i} = \sum_{i=1}^{q} \frac{1}{\lambda_i} \left[\sum_{j=1}^{p} w_{ji} \sqrt{C_{jj}}\ r_{y,x_j} \right]^2 \\ &= \frac{1}{c_{yy}} \sum_{i=1}^{q} \frac{1}{\lambda_i} \left[\sum_{j=1}^{p} w_{ji}\, s_j \right]^2 \end{aligned} \tag{4.5}$$

A. FIRST SCREENING METHOD (M1)

Our goal is to select the best subset of size q from $(x_1,\ldots,x_p)$, where maximization of R^2 is the criterion. The best q principle components are already in hand from (4.4). It seems reasonable to try to march this vector with the "closest" q-dimensional flat in x-space. First we will introduce some notation: the new rotation to the subset of q principle components may be denoted

$$\begin{bmatrix} v_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ v_q \end{bmatrix} = \begin{bmatrix} w_{11}, & \cdots & w_{p1} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ w_{1q}, & \cdots & w_{pq} \end{bmatrix} \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_p \end{bmatrix} . \qquad (4.6)$$

That is,

$$\underset{qx1}{v} = \underset{qxp}{W^{*\prime}} \; \underset{px1}{x} \qquad (4.7)$$

Note that W* consists of q eigenvectors, each of which is complete (that is, the deletion is among the eigenvectors, not across them).

We now have q of the principle components represented as linear combinations of all p of the $\{x_i\}$. The second step is to reduce the number of x_i required. Note that the act of dropping some of the x_i may be viewed as the removal of the corresponding columns of W*' and their replacement with columns of zeroes. Thus

$$\underset{qx1}{v^{**}} = \underset{qxp}{W^{**\prime}} \; \underset{px1}{x} \; . \qquad (4.8)$$

Finding the "closest" subset mentioned above will be accomplished by

$$\min E \, ||v^*-v^{**}||^2 \tag{4.9}$$

where the operator E refers to averaging. Such a minimization must take place for each q=1,...,p. The solution to this problem will be referred to as method M1.

In terms of the elementary quantities, (4.9) may be written as

$$E \, ||v^*-v^{**}||^2 = E \, ||(W^{*\prime}-W^{**\prime})x||^2$$

$$= E \sum_{i=1}^{q} \left[\sum_{j=1}^{p} (w^*_{ji}-w^{**}_{ji})x_j \right]^2 \tag{4.10}$$

$$= \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{p} (w^*_{ji}-w^{**}_{ji})(w_{ki}-w^{**}_{ki})E(x_j x_k)^2.$$

In order to describe the minimization process of (4.9), let S be a set of q subscripts of the variables considered for inclusion in the regression, so that its complement, $\bar{S}$, contains subscripts of those variables to be excluded. Then (4.10) becomes

$$E \, ||v^*-v^{**}||^2 = \sum_{i=1}^{q} \sum_{j\epsilon\bar{S}} \sum_{k\epsilon\bar{S}} w_{ji}w_{ki}C_{jk}. \tag{4.11}$$

Further, let $Q = (Q_1,\ldots,Q_p)$ where

$$Q_i = \begin{cases} 1 \text{ if } i \in S \\ 0 \text{ if } i \notin S \end{cases} . \tag{4.12}$$

Finally let Z_{pxp} be defined

$$(Z)_{kj} = \sum_{i=1}^{q} w_{ji} w_{ki} \cdot \tag{4.13}$$

It follows that (4.10) can be represented as

$$E||v^*-v^{**}||^2 = \sum_{j\epsilon\bar{S}} \sum_{k\epsilon\bar{S}} Z_{jk}C_{jk} = \sum_{j=1}^{p} \sum_{k=1}^{p} Z_{jk}C_{jk}(1-Q_j)(1-Q_k). \tag{4.14}$$

The closest v^{**} may be bound by forming all $\binom{p}{q}$ subsets S of size q, then computing (4.14) for each, and choosing the smallest.

It is useful to express this algorithm in another way. From (4.10) we obtain

$$\begin{aligned} E||v^*-v^{**}||^2 &= \sum_{i=1}^{q} \sum_{j=1}^{p} \sum_{k=1}^{p} (W^{*\prime}-W^{**\prime})_{ij}(W^{*\prime}-W^{**\prime})_{ik}C_{jk} \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} ((W^{*\prime}-W^{**\prime})'(W^{*\prime}-W^{**\prime}))_{jk}C_{jk}. \end{aligned} \tag{4.15}$$

Since

$$(W^{**\prime})_{ij} = (W^{*\prime})_{ij}Q_j \tag{4.16}$$

we see that

$$(W^{*\prime}-W^{**\prime})_{ij} = \begin{cases} 0 & \text{if } x_j \text{ is to be included} \\ w_{ij} & \text{if } x_j \text{ is to be excluded} \end{cases} \tag{4.17}$$

yielding a form that is easily computerized.

B. SECOND SCREENING METHOD (M2)

The rationale for the second method begins with the observation that when (4.5) is used to calculate R^2_{max} (at q=p),

$$R^2_{max} = \frac{1}{c_{yy}} \sum_{i=1}^{p} \frac{1}{\lambda_i} \left[\sum_{j=1}^{p} w_{ji} s_j \right]^2 \tag{4.18}$$

one can decompose the inner summation into two parts; one associated with the variables to be kept and one with the variables to be deleted. The former is indexed by the set S and the latter by its complement $\bar{S}$. Letting

$$t_{ij} = \frac{1}{\sqrt{c_{yy}\lambda_i}} w_{ji} s_j \qquad i,j=1,\ldots,p \tag{4.19}$$

expression (4.18) can be written

$$\begin{aligned} R^2_{max} &= \sum_{i=1}^{p} \left[\sum_{j\in S} t_{ij} + \sum_{j\in \bar{S}} t_{ij} \right]^2 \\ &= R^{*2} + \sum_{i=1}^{p} \{ (\sum_{S} t_{ij})(\sum_{\bar{S}} t_{ij}) + (\sum_{\bar{S}} t_{ij})^2 \}. \end{aligned} \tag{4.20}$$

This serves to define R^{*2} for each set S of indices of variables to be kept. It seems reasonable that a set S that maximizes R^{*2} should define a good set of variables to retain. The process of determining the set S will be referred to as method M2.

The maximization problem may be couched nicely in mathematical programming notation as follows:

$$\begin{aligned} &\text{maximize } R^{*2} = \sum_{i=1}^{p} \left[\sum_{j=1}^{p} t_{ij} Q_j \right]^2 \\ &\text{subject to } \sum_{j=1}^{p} Q_j = q. \end{aligned} \tag{4.21}$$

The expression also lends itself well to computerization.

As in method M1, we must generate all possible Q vectors and compute all possible candidates for (4.21), ultimately choosing the largest for each value of q.

It is interesting that this method can be developed from another point of view. From (2.7) the square of the

multiple correlation may be expressed as

$$R^2_{max} = \frac{1}{c_{yy}} s'C^{-1}s = \frac{1}{c_{yy}} \text{Trace } [ss'C^{-1}]. \qquad (4.22)$$

In hopes of finding a viable screening method, we again consider a subset S of subscripts, and limit the above trace computation to elements of S. To be more explicit, let

$$D = \text{Diag}(Q_1,\ldots,Q_p). \qquad (4.23)$$

Then let us consider screening the vector s with D by looking at the pxp matrix Dss'D. We may think of this as a matrix which has been stamped with a grid so that non-zero elements can be found only in those rows and columns both of whose indices belong to S. Comparison of all possible "stamps" of order q allows us to choose that combination which maximizes the trace in (4.22) Equation (4.22) becomes [Ref. 9]

$$R^{*2} = \frac{1}{c_{yy}} \text{Trace}(Dss'DC^{-1}) = \frac{1}{c_{yy}} \text{Trace}(ss'DC^{-1}D). \qquad (4.24)$$

However, we can show that this is equivalent to method M2 as follows. Let

$$\begin{aligned} \Lambda_{pxp} &= E(vv') = E(W'xx'W) = W'E(xx')W = W'CW \\ &= \text{Diag}(\lambda_1,\ldots,\lambda_p) \end{aligned} \qquad (4.25)$$

and notice that

$$C = W\Lambda W' \quad \text{and} \quad C^{-1} = W\Lambda^{-1}W' \qquad (4.26)$$

properties that follow from the orthogonality of W. It follows that

$$\text{Trace}[ss'DC^{-1}D] = \text{Trace}[ss'DW\Lambda^{-1}W'D]. \qquad (4.27)$$

Since

$$(W\Lambda^{-1}W')_{rj} = \sum_{k=1}^{p} w_{rk}w_{jk}/\lambda_k \tag{4.28}$$

it follows that

$$(DW\Lambda^{-1}W'D)_{rj} = \sum_{k=1}^{p} w_{rk}w_{jk}Q_rQ_j/\lambda_k. \tag{4.29}$$

Since $(ss')_{ir} = s_i s_r$, (4.27) becomes

$$\begin{aligned}\text{Trace}[ss'DW^{-1}W'D] &= \sum_{r=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} s_j s_r w_{rk}w_{jk}Q_rQ_j/\lambda_k \\ &= c_{yy}\sum_{r=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} t_{kj}t_{kr}Q_rQ_j = c_{yy}\sum_{k=1}^{p}\left[\sum_{j=1}^{p} t_{kj}Q_j\right]^2\end{aligned} \tag{4.30}$$

using (4.19). Comparison of this with (4.20) completes the proof. Hence method M2 may be thought of as an attempt to approximate the inverse of a qxq minor of C by a stamped matrix $DC^{-1}D$.

V. RESULTS AND CONCLUSIONS

A. RESULTS

As was mentioned previously, the multiple correlation coefficient is a convenient criterion for judging the effectiveness of the combinations of variables selected by the screening methods under discussion. Thus the results of the screening methods can best be summarized with the graphs of multiple correlation versus the number of variables.

The first test matrix, as mentioned, was designed to expose a weakness in stepwise regression. Figure 1 indicates this quite well. The printout of the SPSS regression program may be reviewed in Table IV. Variable x_1 was selected first because it was highly correlated with y. Stepwise regression then chose variable x_3. Variable x_1 remained significant, and was left in the equation. Stepwise regression then selected variable x_2, found it significant, and terminated. It never observed the pair x_2,x_3 alone, which in this case was a much better pair than the one stepwise regression chose at step two (x_1,x_3). Further, had stepwise regression chosen x_2,x_3, it would not then have selected x_1. This can be seen in Table IV. The reason for this is as follows. The square of the multiple correlation (R^2) of variables x_1,x_3 is a great deal smaller than R^2 for x_2,x_3. Thus when stepwise regression considered adding x_2 to the set x_1,x_3, the sizable increase in R^2 caused it to accept the new variable. On the other hand, when it was given the

CORRELATION COEFFICIENTS

A VALUE OF 99.00000 IS PRINTED
IF A COEFFICIENT CANNOT BE COMPUTED.

	Y	X1	X2	X3
Y	1.00000	0.60000	0.50000	0.50000
X1	0.60000	1.00000	0.50000	0.50000
X2	0.50000	0.50000	1.00000	0.10000
X3	0.50000	0.50000	0.10000	1.00000

DEPENDENT VARIABLE.. Y

VARIABLE(S) ENTERED ON STEP NUMBER 1.. X3
X1
X2

MULTIPLE R	0.70238
R SQUARE	0.49333
STANDARD ERROR	0.73465

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
X3	0.33333	0.33333	0.12368	7.263
X1	0.26667	0.26667	0.14210	3.522
X2	0.33333	0.33333	0.12368	7.263
(CONSTANT)	0.33333			

ALL VARIABLES ARE IN THE EQUATION

DEPENDENT VARIABLE.. Y

VARIABLE(S) ENTERED ON STEP NUMBER 1.. X1

MULTIPLE R	0.60000
R SQUARE	0.36000
STANDARD ERROR	0.80829

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
X1	0.60000	0.60000	0.11547	27.000
(CONSTANT)	2.00000			

Table IV.

* *

VARIABLE(S) ENTERED ON STEP NUMBER 2.. X3

MULTIPLE R 0.64291
R SQUARE 0.41333
STANDARD ERROR 0.78207

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
X1	0.46667	0.46667	0.12901	13.085
X3	0.26667	0.26667	0.12901	4.273
(CONSTANT)	1.33333			

DEPENDENT VARIABLE.. Y

VARIABLE(S) ENTERED ON STEP NUMBER 1.. X3
X2

MULTIPLE R 0.67420
R SQUARE 0.45455
STANDARD ERROR 0.75410

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
X3	0.45455	0.45455	0.10827	17.625
X2	0.45455	0.45455	0.10827	17.625
(CONSTANT)	0.45455			

* *

VARIABLE(S) ENTERED ON STEP NUMBER 2.. X1

MULTIPLE R 0.70238
R SQUARE 0.49333
STANDARD ERROR 0.73465

---------- VARIABLES IN THE EQUATION ----------

VARIABLE	B	BETA	STD ERROR B	F
X3	0.33333	0.33333	0.12368	7.263
X2	0.33333	0.33333	0.12368	7.263
X1	0.26667	0.26667	0.14210	3.522
(CONSTANT)	0.33333			

ALL VARIABLES ARE IN THE EQUATION

Table IV. continued.

DEPENDENT VARIABLE.. Y

VARIABLE(S) ENTERED ON STEP NUMBER 3.. X2

MULTIPLE R	0.70238
R SQUARE	0.49333
STANDARD ERROR	0.73465

VARIABLES IN THE EQUATION

VARIABLE	B	BETA	STD ERROR B	F
X1	0.26667	0.26667	0.14210	3.522
X3	0.33333	0.33333	0.12368	7.263
X2	0.33333	0.33333	0.12368	7.263
(CONSTANT)	0.33333			

MAXIMUM STEP REACHED

Table IV. continued.

opportunity to add x_1 to the pair x_2,x_3, there was not sufficient improvement, and x_1 was not added. Since stepwise regression missed the best pair, it did not terminate until all three were included, whereas it could have terminated with two variables had it found the best pair.

Note also that while method M1 falls into the same trap as stepwise regression, method M2 selected the same combinations as did total enumeration for each q=1,2,3. The reader will also observe that the principle component multiple correlation curve serves quite well in this graph, and the two that follow, as a standard with which the other methods may be compared.

It is interesting that this curve and the total enumeration curve are generally very close in the cases presented here, and in fact cross each other in the second case. This observation is useful since in many applications the total enumeration solution is unavailable.

Figure 2 indicates the results of the various methods with the second test matrix. Nore that while the R^2 curve for method M2 runs well with the total enumeration and stepwise regression curves (identical curves in this case), the curve for method M1 runs consistently lower throughout the entire midrange.

Figure 3 presents the results of the screening methods on the third test matrix. This is of particular interest because the results should be useful in the study cited earlier. Note again that the curve for M2 runs quite strongly with the curve for stepwise regression, but that the curve

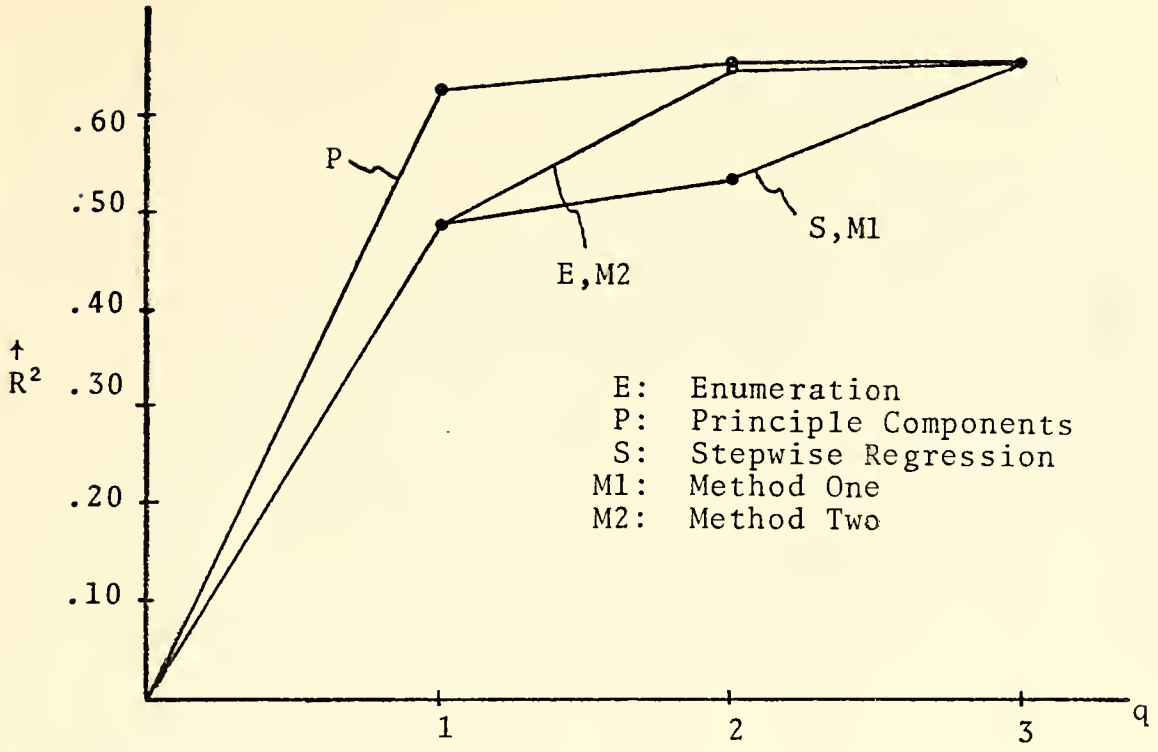


Figure 1. Test Matrix One Results.

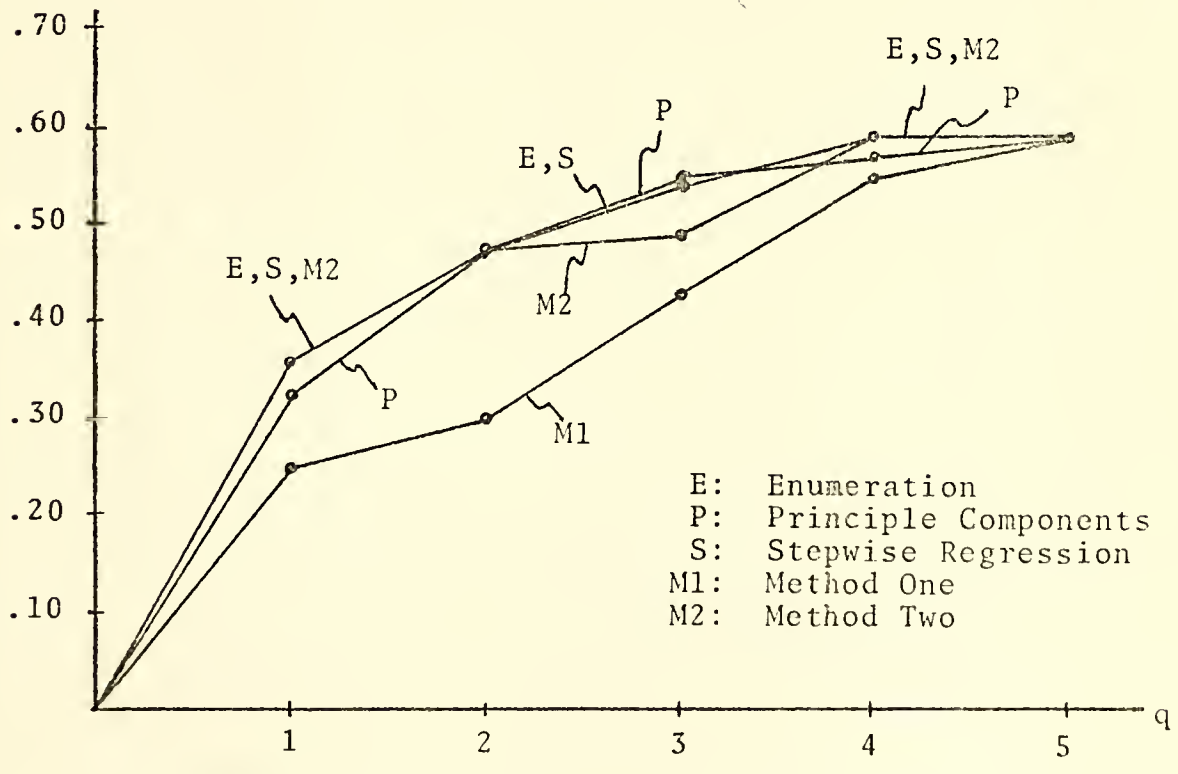


Figure 2. Test Matrix Two Results.

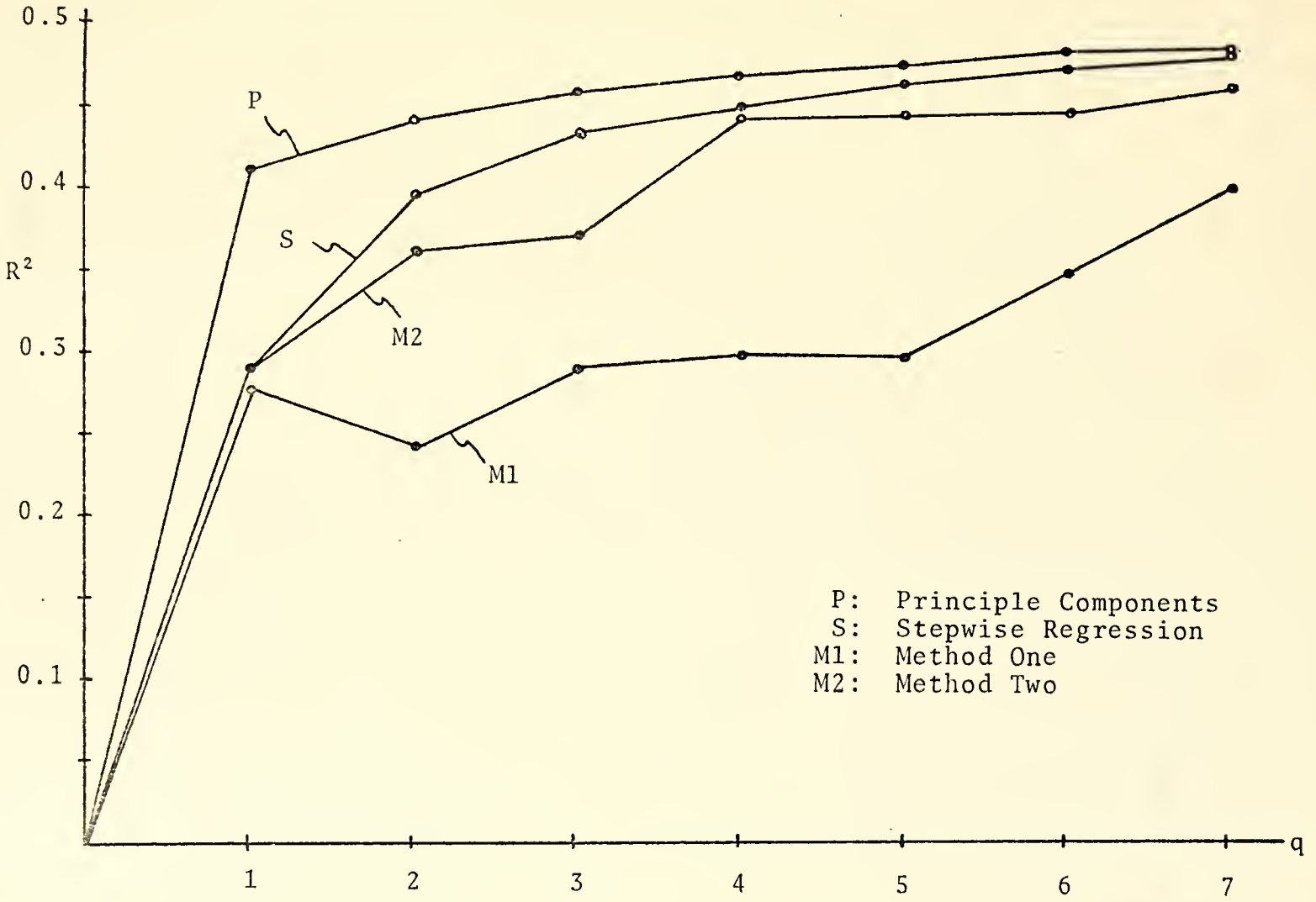


Figure 3. Results of Test Matrix Three.

for M1 falls consistently below the others. Throughout the midrange of all three test cases, the curve for M1 is about two-thirds of the total enumeration curve.

Because the second method shows greater promise as a screening device, more detailed results on its performance will be presented. Unlike stepwise regression, this method is capable of ordering all combinations of size q according to their R^{*2} values. (Stepwise regression will typically only look at one or two of the combinations.) Thus, Table V contains a summary of the combinations selected by the second method, M2.

From this table, the graph in Figure 4 was constructed, giving R^{*2} as a function of the number of variables. This curve seems fairly typical of what can be expected. Note that as the number of variables allowed to enter increases, R^{*2} initially increases. At some point the curve will peak, then decrease to the multiple correlation value of the entire set of variables. This phenomenon can be explained in the following way. The value of T_{ij} (see 4.19, 4.20) may be of either sign. When the number of variables is small, it is fairly likely that some combination of variables will be such that the t_{ij}'s will combine with the same sign, so that when squared and summed again the value may get quite large. (In the three examples used in this paper, the value typically exceeded one.) However, as the number of variables increases, it becomes much more likely that the t_{ij}'s, differing in sign, will conflict with each other and reduce the overall values being squared

Q=1 Vbles	R^{*2}	Q=2 Vbles	R^{*2}	Q=3 Vbles	R^{*2}	Q=4 Vbles	R^{*2}
x_1	.584	x_1x_2	.524	$x_1x_3x_4$	1.13	$x_1x_2x_3x_4$	1.002
x_2	.015	x_1x_3	.965	$x_1x_2x_3$	.871	$x_1x_3x_4x_5$	.676
x_3	.215	x_1x_4	.741	$x_1x_2x_4$	.650	$x_1x_2x_3x_5$	.565
x_4	.115	x_1x_5	.498	$x_1x_3x_5$	.627	$x_1x_2x_4x_5$	.426
x_5	.430	x_2x_3	.196	$x_1x_4x_5$	.550	$x_2x_3x_4x_5$	.391
		x_2x_4	.100	$x_2x_4x_5$	.518		
		x_2x_5	.478	$x_1x_2x_5$	.472		
		x_3x_4	.333	$x_2x_3x_5$	.406		
		x_3x_5	.393	$x_3x_4x_5$	.351		
		x_4x_5	.440	$x_2x_3x_4$	.296		

Table V. Combinations Chosen by Method 2.

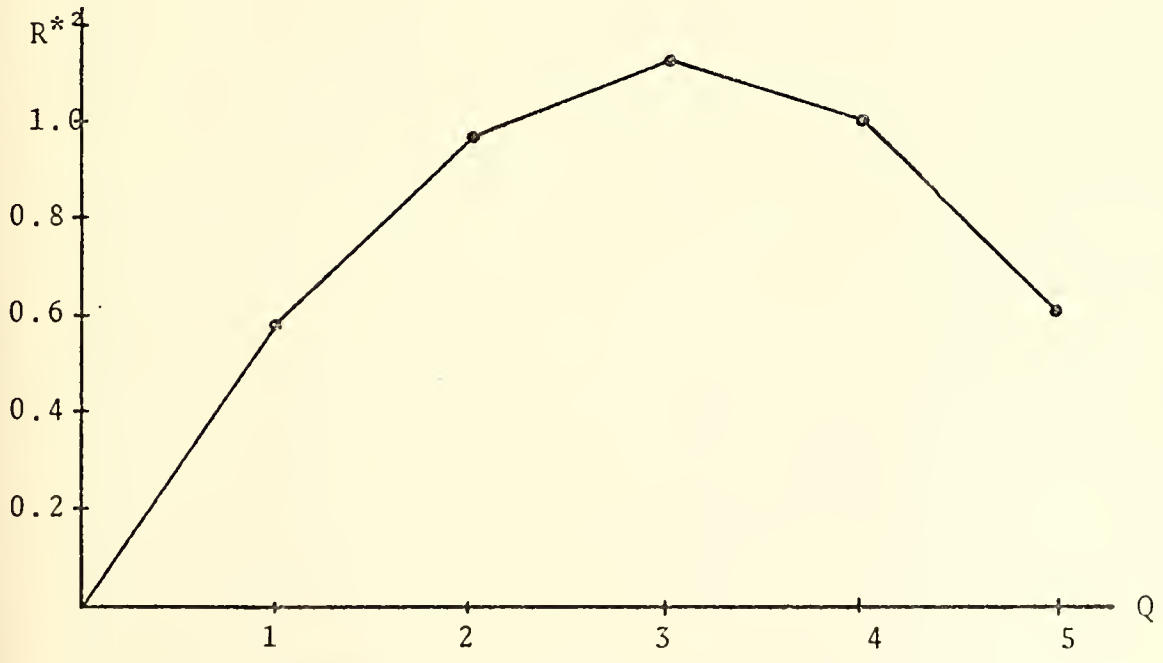


Figure 4. Graph of R^{*2} vs. Q for Method 2.

and summed. The curve in Figure 4 seems typical of what may be expected. Thus at best these values are pseudo-correlations, which have been shown to rank well with, but not predict, the ranking of the actual multiple correlation coefficients. As a result, this method cannot guarantee optimal selections.

Table VI gives the rankings of total enumeration and of method M2 for each combination of size q=3 and q=4 from test case two. In this way, the results of the method may be more fully evaluated than to consider only the "best" selection made at each step. Using the rankings given in the table, Spearman's Rank Correlation test [Ref. 10] was applied; at α=.1 the rankings of total enumeration were significantly correlated with the rankings of M2 in both cases. This lends credence to the second screening method, and also indicates a convenient property of this method; the capability to rank all combinations at each step.

Finally some comments on the use of principle components are in order. It appears to provide a useful standard of comparison for the other methods presented. Indeed, the method may be of more direct use in some applications, especially when researchers find that $p \geq N$. In such cases the coefficients β may not be directly estimable, but any version of least squares will result in $\hat{\varepsilon}=0$. Principle components may be useful for preliminary screening so that there are enough degrees of freedom N-q to obtain an acceptable estimate of variance of residuals.

Q=3	Princ. Components		Multiple Correlation	
Variables	R^{*2}	Rank	R^2	Rank
$x_1x_3x_4$	1.126	1	.492	2
$x_1x_2x_3$	.871	2	.545	1
$x_1x_2x_4$	.650	3	.448	4
$x_1x_3x_5$	.627	4	.487	3
$x_1x_4x_5$	.550	5	.431	6
$x_2x_4x_5$	.518	6	.278	9
$x_1x_2x_5$	.472	7	.437	5
$x_2x_3x_5$	.406	8	.301	8
$x_3x_4x_5$	.351	9	.317	7
$x_2x_3x_4$	.296	10	.223	10
Q=4	Spearman's ρ = 0.93			
$x_1x_2x_3x_4$	1.000	1	.593	1
$x_1x_3x_4x_5$	.676	2	.498	3
$x_1x_2x_3x_5$	.565	3	.549	2
$x_1x_2x_4x_5$	.426	4	.478	4
$x_2x_3x_4x_5$	.391	5	.329	5

Spearman's ρ = 0.80

Table VI.

Compared Rankings of the Second Method With Enumeration

B. CONCLUSIONS

On the basis of the results just presented, a number of observations concerning the screening methods under investigation in this paper may be made.

The first method (M1) does not appear to perform particularly well. Its selections were consistently worse than those made by the other methods. It may be that its disappointing performance is caused by the weak correlation of the major principle components with the response variable. As was seen in the three examples presented, this method does not in general even approach the optimal combinations, and thus shows little promise per se as a screening technique.

Method M2 seems to be a reasonable approach to the problem. While it will not in general make optimal selections, in the examples used here it has done quite well.

The method has several advantages which are worthy of mention. The first is that after an initial investment in obtaining the eigenvalues and eigenvectors of C (roughly equivalent to inverting it, in time spent), the amount of time required to examine any potential combination is very small. As a result of this, enumeration of the combinations becomes feasible even when the number of variables is fairly large. Appendix C contains some remarks concerning an algorithm which will enumerate quite efficiently, and which was used in the FORTRAN program presented there. This program outputs the largest value of R^{*2}, and the combination that produced it, for each value of q.

It is worth noting that the manner in which M2 screens variables is much more readily apparent than that of stepwise regression. The user is aware of the process by which variables are screened, and is more able to make intelligent use of it.

C. SUMMARY

The results presented here are at best tentative, since only three test cases are presented. Certainly a wider spectrum of test cases is necessary to establish conclusive results for the methods presented in this paper. Neither stepwise regression nor method M2 is optimal, but both remain as viable competitors which are useful as screening devices.

Because of the success of M2, we are led to consider the possibility of other methods of approximating the inverses of the minors of C, which may be as efficient (with computer resources), as those presented yet produce results which compare more favorably with total enumeration.

APPENDIX A: PRINCIPLE COMPONENTS

The structure for principle components [Ref. 11] arose from a desire to find that location and orientation of axes that the variance of the data swarm about them is minimized. The necessary location of axes follows from a linear algebra theorem which states that a sum of squares centered about an arbitrary point is minimized when that point is the centroid. Thus it becomes convenient to standardize the data about their means. In order to define the desired orientation, let us first agree to the following conventions.

Let the data swarm under consideration be in p-space. Let the X matrix contain the p coordinates of the N observations, and let

$$\bar{x}_{.j} = \frac{1}{N} \sum_{i=1}^{N} x_{ij} \quad \text{and} \quad s_j^2 = \frac{1}{N} \sum_{i=1}^{N} (x_{ij} - \bar{x}_{.j})^2. \tag{A.1}$$

Then let the X matrix have entries

$$(X)_{ij} = ((x_{ij} - x_{.j})/s_j). \tag{A.2}$$

That is, let the X matrix contain the p coordinates of the N observations after the variates have been standardized to zero mean and unit variance. (Note that unit variance is not necessary to what follows.) Then it follows that the correlation matrix is

$$C_{pxp} = \frac{1}{N} X'X. \tag{A.3}$$

The rotation itself may be denoted

$$\underset{p\times 1}{v} = W'x \tag{A.4}$$

where $v = (v_1, v_2, \ldots, v_p)'$, $x = (x_1, x_2, \ldots, x_p)'$, and W_{pxp} is a matrix column-oriented vectors, each of which provides the coefficients for transforming the x vector into one of the component directions.

For p=2, this may be displayed graphically (see Figure 5).

As stated previously, the goal is to minimize the variance about the component axes by choosing the matrix W. However, one may do this neatly by setting $W=\underline{0}$, and thus avoid the real problem. To obtain a tenable solution we may constrain the problem by requiring that the norm is one: $w'w=1$. The variance about the component directions may be written

$$\begin{aligned} S^2 &= E(w'x)^2 = E(w'xx'w) = w'E(cc')w \\ &= w'Cw. \end{aligned} \tag{A.5}$$

Because minimizing the variance about an axis is equivalent to maximizing the variance along that axis, and because S^2 represents the variance along the axes v_1, our problem will be to maximize S^2. To show that this is true, refer to Figure 6. Note that to minimize the variance in the v_2 direction we must maximize along the (orthogonal) v_1 direction. Orthogonality will be shown later.

Thus the problem of finding the direction V_1 may be written

$$\text{maximize } w'Cw \quad \text{subject to } w'w = 1 \tag{A.6}$$

where w is an arbitrary column of W. This problem may be solved conveniently by the use of LaGrange multipliers. The

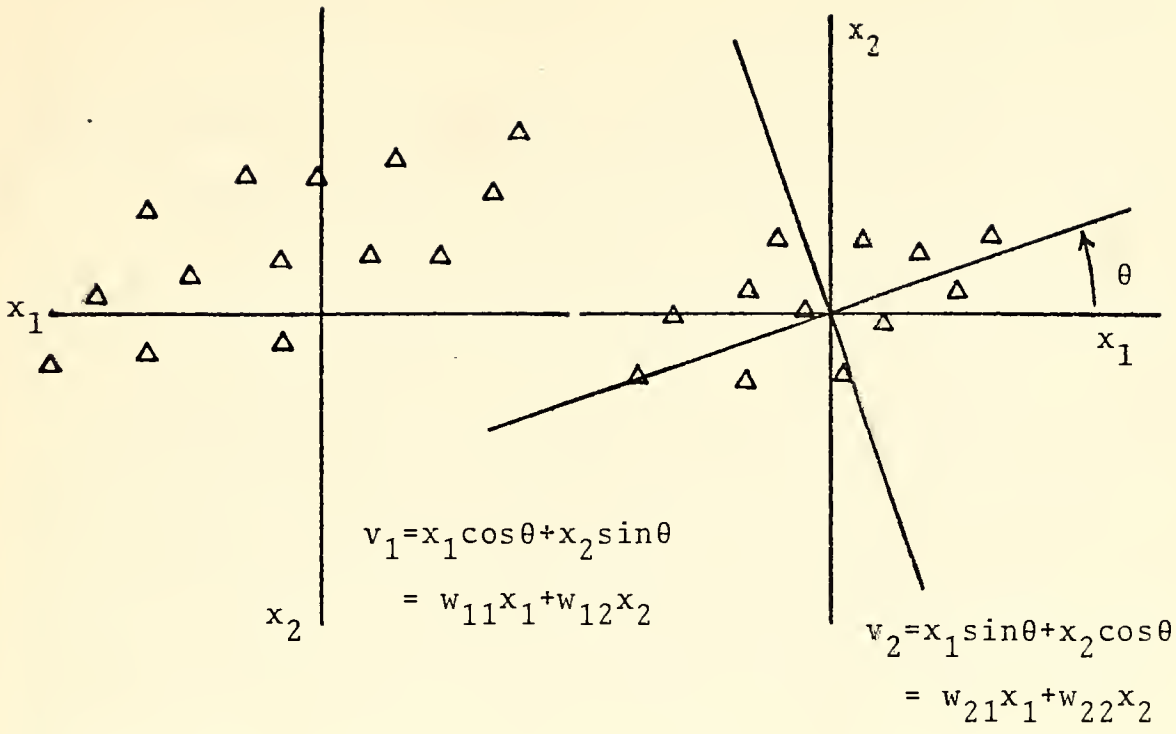


Figure 5. Principle Component Rotation.

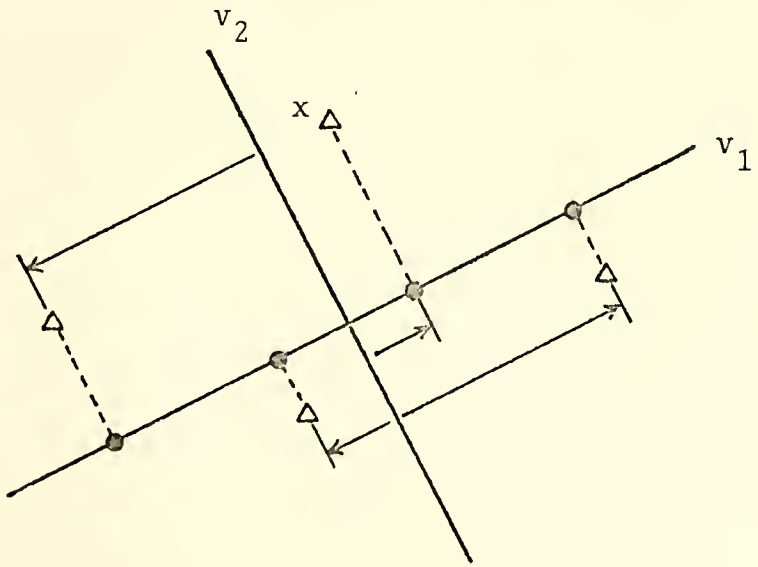


Figure 6. Maximizing the Variance Along v_1.

LaGrangian is

$$\phi_1 = w'Cw - \lambda(w'w-1). \tag{A.7}$$

The LaGrangian is maximized at $\nabla\phi_1=0$

$$\nabla\phi_1 = 2(C-\lambda I)w = 0. \tag{A.8}$$

Then

$$(C-\lambda I) = 0. \tag{A.9}$$

Now the matrix C is real and symmetric, and will in general be positive definite; it can be shown to be at least positive semi-definite as follows: Let Z be a non-negative vector, so that $Z'CZ = Z'X'XZ = (XZ)'XZ$. Now let $Y = XZ$, so that $(XZ)'XZ = Y'Y = \sum_{i=1}^{N} y_i^2 \geq 0$ for all z_i in the vector Z.

As a result, it can be shown [Ref. 9] that C is non-singular with real, non-negative eigenvalues.

We require that $(C-\lambda I)=0$. If $(C-\lambda I)$ is non-singular, the only solution to the simultaneous equations is $w=\underline{0}$, which violates the constraint $w'w=1$. Thus we require that $(C-\lambda I)$ be singular, and it follows that

$$|C-\lambda I| = 0. \tag{A.10}$$

Then the set of p solutions λ_i to this equation must be the characteristic values of the matrix C. Pre-multiplying equation A.9 by w' we obtain

$$w'(C-\lambda I)w = w'\cdot 0 = 0. \tag{A.11}$$

Then

$$w'Cw = w'\lambda Iw = \lambda w'w = \lambda. \tag{A.12}$$

But $w'Cw = S^2$. Thus λ is actually the variance along the component. Since we wish to maximize the variance, the solution we desire is the largest eigenvalue.

Because λ is an eigenvalue, (A.9) implies that w is its associated eigenvector, and thus we maximize S^2 by choosing the eigenvector associated with the largest eigenvalue as the direction v_1. Then call this eigenvector w_1 and its eigenvalue λ_1.

The direction v_2 may be found solving the following equation:

$$\text{maximize } w_k'Cw_k \text{ subject to } w_k'w_k = 1 \text{ and } w_1'w_k = 0. \quad \text{(A.13)}$$

The last constraint says that we require v_1 and v_2 to be orthogonal. The LaGrangian is

$$\phi_2 = w_k'Cw_k - \theta(w_k'w_k-1) - 2\tau(w_1'w_k-0) \quad \text{(A.14)}$$

and

$$\nabla\phi_2 = 2Cw_k - 2\theta w_k - 2\tau w_1 = 0. \quad \text{(A.15)}$$

Then

$$(C-\theta I)w_k - \tau w_1 = 0. \quad \text{(A.16)}$$

Pre-multiplying by w_k',

$$\begin{aligned} &w_k'(C-\theta I)w_k - w_k'w_1 = 0 \rightarrow w_k'(C-\theta I) = 0 \rightarrow \\ &w_k'Cw_k = \theta w_k'Iw_k = \theta w_k'w_k = \theta \rightarrow S^2 = \theta. \end{aligned} \quad \text{(A.17)}$$

Thus θ is one of the eigenvalues of the C matrix, and w_k its associated eigenvector. Now pre-multiply (A.16) by w_1',

$$\begin{aligned} &w_1'(C-\theta I)w_k - \tau w_1'w_1 = 0 \rightarrow w_1'(C- I)w_k = \tau \rightarrow \\ &w_1'Cw_k - \theta w_1' = \tau = w_1'Cw_k . \end{aligned} \quad \text{(A.18)}$$

Pre-multiplying (A.9) by w_k',

$$w_k'(C-\lambda I)w_1 = w_k' \cdot 0 = 0 \rightarrow w_k'Cw_1 = \lambda w_k'w_1 \rightarrow w_k'Cw_1 = 0. \quad (A.19)$$

Then.

$$w_k'Cw_1 = \tau = 0. \quad (A.20)$$

Note that because τ turned out to be zero, the constraint that forced the new component to be orthogonal to the first component was not binding; that is, w_k is inherently orthogonal to w_1. This follows because C is a real, symmetric, positive-definite matrix. Since $\tau=0$, (A.16) becomes identical in form to (A.9), so that θ is an eigenvalue, w_k its associated eigenvector. It was shown above that $\theta=S_2^2$ (θ is the variance in the v_2 direction), so that it follows naturally that θ is the second-largest eigenvalue (let $\theta=\lambda_2$). Then w_k becomes w_2.

This argument can be generalized so that λ_i, the i^{th} largest eigenvalue of C, is the variance in the i^{th} best component direction v_i, and the associated eigenvector w_i is the set of coefficients mapping the x vector into V_i. Now let Λ_{pxp} be a diagonal matrix with the ordered eigenvalues (largest to smallest) in the diagonal. The rotation we desire is then

$$v = W'x,$$

and the correlation matrix of the v_i is Λ.

APPENDIX B: PROPERTIES OF THE CORRELATION MATRIX

Throughout the course of this paper, and especially in the numerical examples, there has been an implicit assumption that if a matrix is real, symmetric, and positive definite, it qualifies as a correlation matrix, and that it would be possible to find real data that would generate such a correlation matrix. While this is not overly difficult to show, it is of general interest, and is not contained in many textbooks as a complete derivation.

The mathematical statement of the problem is to find a matrix X_{Nxp} such that $X'X/N = C_{pxp}$, where C is given. It has been shown previously that to qualify as the product of a matrix and its transpose, C must be positive semi-definite. (If C has full rank, it will be a positive definite matrix.) Further, by the definition of correlation between two variables, C must be real and symmetric.

To find the X matrix, the initial step is to find a triangular matrix T_{pxp} such that $T'T = C$. It is possible to compute the elements of the T matrix by carrying out the multiplication (shown in this case for a 3x3): Let T be upper triangular, so that

$$\begin{bmatrix} T_{11} & 0 & 0 \\ T_{12} & T_{22} & 0 \\ T_{13} & T_{23} & T_{33} \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} & T_{13} \\ 0 & T_{22} & T_{23} \\ 0 & 0 & T_{33} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12} & C_{22} & C_{23} \\ C_{13} & C_{23} & C_{33} \end{bmatrix}$$

or

$$\begin{bmatrix} T_{11}T_{11} & T_{11}T_{12} & T_{11}T_{13} \\ T_{11}T_{12} & T_{12}T_{12}+T_{22}T_{22} & T_{12}T_{13}+T_{22}T_{23} \\ T_{11}T_{13} & T_{13}T_{12}+T_{23}T_{22} & T_{13}T_{13}+T_{23}T_{23}+T_{33}T_{33} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12} & C_{22} & C_{23} \\ C_{13} & C_{23} & C_{33} \end{bmatrix}$$

Thus

$$T_{11} = \sqrt{C_{11}} \qquad T_{12} = C_{12} / \sqrt{C_{11}} \qquad T_{13} = C_{13} / \sqrt{C_{11}}$$

$$T_{22} = \sqrt{C_{22} - C_{12}^2 / C_{11}} \qquad T_{23} = \frac{C_{23} - C_{12}C_{13} / C_{11}}{\sqrt{C_{22} - C_{12}^2 / C_{11}}}$$

$$T_{33} = \sqrt{C_{33} - \frac{C_{13}^2}{C_{11}} - \frac{(C_{23} - C_{12}C_{13} / C_{11})^2}{C_{22} - C_{12}^2 / C_{11}}}$$

Note that at every step it is possible to solve for T_{ij} in terms of the C_{ij} and only those other T_{ij} for which it has already been possible to solve. This may be done in general for a C matrix of any size p.

Working backwards, we have that

$$T'T = C = X'X/N. \tag{B.1}$$

Then let

$$\underset{Nxp}{Z} = \underset{Nxp}{X} \; \underset{pxp}{T^{-1}} \; . \tag{B.2}$$

Then

$$Z'Z = (T^{-1})'X'XT^{-1} = (T^{-1})CT^{-1} = (T^{-1})T'TT^{-1} = I \; . \tag{B.3}$$

Let $Z \sim N(\underline{0}, I)$, and we obtain the following transformation:

$$X = ZT. \tag{B.4}$$

Then

$$\mathcal{E}X'X = T'Z'ZT = T'IT = T'T = C.$$

Thus having obtained a matrix Z containg N observations on a p-variate standard normal, it is possible to transform Z to a new set of N observations on a p-variate normal (though no longer with unit variance) which will generate the desired correlation matrix through the judicious choice of a triangular matrix T. Thus the only requirements on a matrix are that it be real, symmetric, and positive-definite in order to be a correlation matrix [Ref. 9].

APPENDIX C: FORTRAN PROGRAM OF SECOND SCREENING METHOD FOR APPLICATION IN TEST CASE THREE

Because of the generally good results given by the second screening method suggested in this paper, the FORTRAN program which was used to apply it to the third matrix is listed here. The program does the following:

1. It reads a correlation matrix (then reverses the last two rows and columns, as in this case X13 is to be the response variable) and also reads a standard deviation vector. From these it obtains the covariance matrix. (Lines 1 to 40, beginning the count with the dimension statement.)

2. It calls a subroutine (JACVAT) which calculates the eigenvalues and eigenvectors of the covariance matrix. (Lines 34 to 50).

3. The equation developed under the first approach to the second screening method is used to calculate the multiple correlation coefficient. The printout, under the label "R-square equals," and showing the sums and squares, may be used to apply the screening method by hand. (Lines 50-73.)

4. The covariance matrix's inverse is calculated for transfer to the subroutine USER in order to use it with the second approach to the second method. (Lines 74-90.)

5. The next part of the program is the "driver program" for the subroutine TWIDDL (which provides the vector of ones and zeroes). It provides the initial vector (Q), and performs the one-zero exchange. (Lines 93 to 120).

6. The TWIDDL subroutine is an algorithm originally written in Algol for the Association of Computing Machinery, and is listed as Algorithm 384 in the CACM. Essentially it forms every possible combination of m ones and (n-m) zeroes, but each new vector requires only the interchanging of two positions in the previous vector [Ref. 12].

7. The USER subroutine uses the inverse of the covariance matrix, the indicator vector Q, and the augmented covariance matrix to screen variables using the second approach. For each value of m, it continually stores the largest value of R_S^2 that it has calculated to date, and when m increments it prints the value of R_S^2 and the vector Q which produced it.

(Using this program as a backbone, it is a straightforward matter to obtain a program which will apply the first method to the data matrix.)

```
//LAMBELL JOB (2629,1242,RL42),'LAMBELL',TIME=3
// EXEC FORTCLG,REGION.GO=70K
//FORT.SYSIN DD *
      DIMENSION A(14,14),B(13,13),C(13,13),D(13),SD(14),
     1F(14,14),Q(13)
      COMMON PO,P(13)
      INTEGER HOLDQ(13)
      INTEGER PO,P,X,Y,Z,DONE,Q
      NEW=0
      DO 20 I=1,14
      READ(5,10)(A(I,J),J=1,8)
      READ(5,10)(A(I,J),J=9,14)
   10 FORMAT(8(2X,F8.5))
   20 CONTINUE
      DO 30 J=1,14
      TEMP=A(13,J)
      A(13,J)=A(14,J)
   30 A(14,J)=TEMP
      DO 40 J=1,14
      TEMP=A(J,13)
      A(J,13)=A(J,14)
   40 A(J,14)=TEMP
      WRITE(6,50)
   50 FORMAT(1X,/////,'  THE CORRELATION MATRIX IS')
      DO 70 I=1,14
      WRITE(6,60)(A(I,J),J=1,14)
   60 FORMAT(1X,//,1X,14F9.3)
   70 CONTINUE
      READ(5,80)(SD(I),I=1,8)
      READ(5,80)(SD(I),I=9,14)
   80 FORMAT(8(2X,F8.5))
      DO 90 I=1,14
      DO 90 J=1,14
   90 F(I,J)=A(I,J)*SD(I)*SD(J)
      DO 100 I=1,13
      DO 100 J=1,13
  100 B(I,J)=F(I,J)
      CALL JACVAT(8,13,1,D,C,13)
      WRITE(6,110)
  110 FORMAT(1X,////,'  THE COVARIANCE MATRIX IS')
      DO 130 I=1,14
      WRITE(6,120)(F(I,J),J=1,14)
  120 FORMAT(1X,//,1X,14F9.3)
  130 CONTINUE
      WRITE(6,140)
  140 FORMAT(1X,////,'  THE EIGENVALUES ARE')
      WRITE(6,150)(D(I),I=1,13)
  150 FORMAT(1X,//,1X,13F9.3)
      WRITE(6,160)
  160 FORMAT(1X,///,' EIGENVECTORS ARE IN THE COLUMNS')
      DO 180 I=1,13
      WRITE(6,170)(C(I,J),J=1,13)
  170 FORMAT(1X,//,1X,13F9.5)
  180 CONTINUE
      DO 190 I=1,13
      DO 190 J=1,13
  190 B(I,J)=C(J,I)*A(14,J)*SD(J)/SQRT(D(I))
      WRITE(6,200)
  200 FORMAT(1X,////,'  R-SQUARE EQUALS')
      DO 220 I=1,13
      WRITE(6,210)(B(I,J),J=1,13)
  210 FORMAT(1X,//,'  (',12(F8.5,'+'),F8.5,')**2 +')
  220 CONTINUE
      DO 240 I=1,13
      SUM=0.0
      DO 230 J=1,13
  230 SUM=SUM+B(I,J)
  240 SD(I)=SUM**2
      WRITE(6,250)
  250 FORMAT(1X,////,'  OR R-SQUARE EQUALS')
      WRITE(6,260)(SD(I),I=1,13)
  260 FORMAT(1X,//,1X,12(F9.6,'+'),F9.6)
```

```
      SUM=0.0
      DO 270 I=1,13
270   SUM=SUM+SD(I)
      WRITE(6,280) SUM
280   FORMAT(1X,////,'   OR R-SQUARE EQUALS',F12.9)
      N=13
      DO 400 M=1,6
      PO=N+1
      P(N+1)=-2
      DO 320 I=1,N
      IF(I.GT.N-M) GO TO 310
      P(I)=0
      GO TO 320
310   P(I)=I-N+M
320   CONTINUE
      DONE=0
      DO 340 I=1,N
      IF(I.GT.N-M) GO TO 330
      Q(I)=0
      GO TO 340
330   Q(I)=1
340   CONTINUE
350   CALL USER(B,Q,M,NEW,F,HOLDQ,HOLD)
      CALL TWIDDL(X,Y,Z,DONE)
      IF(DONE.EQ.1) GO TO 370
      Q(X)=1
      Q(Y)=0
      GO TO 350
370   CALL USER(B,Q,M,NEW,F,HOLDQ,HOLD)
400   CONTINUE
      WRITE(6,410) (HOLDQ(I),I=1,13),HOLD
410   FORMAT(1X,'Q=  6',5X,'(',13I2,' )',5X,'R*2=',E16.7)
      STOP
      END
      SUBROUTINE TWIDDL(X,Y,Z,DONE)
      COMMON PO,P(20)
      INTEGER X,Y,Z,DONE,PO,P
      J=0
 10   J=J+1
      IF(P(J).LE.0) GO TO 10
      IF(P(J-1).NE.0) GO TO 40
      I=J-1
 20   IF(I.LT.2) GO TO 30
      P(I)=-1
      I=I-1
      GO TO 20
 30   P(J)=0
      Z=1
      X=Z
      P(1)=X
      Y=J
      RETURN
 40   IF(J.GT.1) P(J-1)=0
 50   J=J+1
      IF(P(J).GT.0) GO TO 50
      K=J-1
      I=K
 60   I=I+1
      IF(P(I).NE.0) GO TO 70
      P(I)=-1
      GO TO 60
 70   IF(P(I).NE.-1) GO TO 80
      Z=P(K)
      P(I)=Z
      X=I
      Y=K
      P(K)=-1
      RETURN
 80   IF(I.EQ.PO) GO TO 90
      P(J)=P(I)
      Z=P(J)
      P(I)=0
```

```
      X=J
        Y=I
      RETURN
90    DONE=1
      RETURN
      END
      SUBROUTINE USER(B,Q,M,NEW,HOLDQ, HOLD)
      DIMENSION B(13,13),C(13,13),Q(13),HOLDQ(13),PSUM(13)
      INTEGER Q,HOLDQ
      DO 10 I=1,13
      DO 10 J=1,13
10    C(I,J)=B(I,J)*Q(J)
      DO 20 I=1,13
      PSUM(I)=0.0
      DO 15 J=1,13
15    PSUM(I)=PSUM(I)+C(I,J)
20    PSUM(I)=PSUM(I)**2
      SUM=0.0
      DO 30 I=1,13
      SUM=SUM+PSUM(I)
30    CONTINUE
      IF(NEW.EQ.0) GO TO 50
      IF(M.NE.NEW) GO TO 80
      NEW=M
      GO TO 60
50    HOLD=NEW
      NEW=M
60    IF(HOLD.GT.SUM) RETURN
      HOLD=SUM
      DO 70 I=1,13
70    HOLDQ(I)=Q(I)
      RETURN
80    L=M-1
      WRITE(6,40) L,(HOLDQ(I),I=1,13),HOLD
40    FORMAT(1X,'Q=',I3,5X,'(',13I2,' )',5X,'R*2=',E16.7)
      HOLD=0.0
      NEW=M
      GO TO 60
      END
```

LIST OF REFERENCES

1. Draper, Norman, R. and S. Smith, Applied Regression Analysis, New York, Wiley, 1966.

2. Anderson, T. W., An Introduction to Multivariate Statistical Analysis, Wiley, 1958, pp.86-96, 272-286.

3. Statistical Package for the Social Sciences, Norman Nie and Dale H. Bent and C. Hadlai Hull, McGraw-Hill, 1970, pp. 174-195.

4. Biomedical Computer Programs, W. J. Dixon, editor, University of California Press, Third Edition 1973, pp. 193-210, 285-352.

5. Morrison, Donald W., Multivariate Statistical Methods, McGraw-Hill, 1967.

6. Hocking, R. R., and R. N. Leslie, 1967, "Selection of the Best Subset in Regression Analysis," Technometrics 9, 531-540.

7. Hoerl, Arthur E., and Robert W. Kennard, 1970, "Ridge Regression: Applications to Nonorthogonal Problems," Technometrics, 12, 69-82.

8. Read, R. R. and T. A. Wyatt, "Some Scaling Methods Applied to the Perception of Drugs," (1975) Naval Postgraduate School Technical Report, unpublished.

9. Graybill, Franklin A., Introduction to Matrices with Applications in Statistics, Wadsworth Publishing Co., 1969, pp. 39-51, 163-235.

10. Mendenhall, William and James E. Reinmuth, Statistics for Management and Economics, Duxbury Press, 1971, pp. 469-472.

11. Harmon, Harry H., Modern Factor Analysis, University of Chicago Press, 1967, pp. 135-186.

12. "Collected Algorithms from CACM," Association for Computing Machinery, Inc., 1970, Algorithm 382, Combinations of M Out of N Objects (G6), Philip J. Chase.

INITIAL DISTRIBUTION LIST

		No. Copies
1.	Defense Documentation Center Cameron Station Alexandria, Virginia 22314	2
2.	Library, Code 0212 Naval Postgraduate School Monterey, California 93940	2
3.	Department Chairman, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	2
4.	Professor Robert R. Read, Code 55 Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
5.	ENS Dennis George Lambell 4405 N. Rosemead Blvd., Apt. #215 Rosemead, California 91770	1
6.	Chief of Naval Personnel Pers 11b Department of the Navy Washington, D. C. 20370	1

Thesis 157298
L2513 Lambell
c.1 Some alternative methods to stepwise regression for the screening of variables.